CoSeC SOFTWARE SPOTLIGHT

USolv

https://cloud.ccp4.ac.uk/

David McDonagh (david.mcdonagh@stfc.ac.uk)

USolv is a machine learning approach to predicting protein crystal solvent content, which can be a key bottleneck in the protein structure solution pipeline. The models are based on the Vit-V-Net transformer and U-Net convolutional neural network architectures, developed using PyTorch in Python. Read more about USolv in the preprint https://doi.org/10.1101/2025.09.24.678396.

Value for Research

X-ray crystallography is the primary experimental technique for understanding the structure of proteins, forming the backbone of research in drug discovery and biotechnology. In these experiments, a protein is purified, crystallised, and placed in an X-ray beam. The subsequent diffraction pattern is then used to determinate the charge density of the protein, which in turn can be used to infer the positions of atoms. Diffraction patterns are typically generated at synchrotrons, such as Diamond Light Source, and data processing requires a suite of software, such as the suite developed by The Collaborative Computational Project No. 4 (CCP4), of which USolv is now a part.

Estimating the solvent content of the crystal is a vital step in this processing as this determines the number of molecules in the repeating unit of the crystal. Getting this wrong can significantly impede the structure solution pipeline, wasting compute resources and reducing the chance of resolving the structure. Traditional approaches in CCP4 fail to correctly predict the number of molecules correctly in around 20% of cases. USolv more than halves this error, significantly improving this bottleneck.

Development

USolv was designed and developed by Dr David McDonagh at the Scientific Computing Department of STFC, as part of CCP4. The project was supervised by Dr Ronan Keegan, with additional input from Dr David Waterman, and Professor Dan Rigden from the University of Liverpool.

Functionality

USolv is deployed as part of CCP4 and works by predicting the solvent content of a structure from a Patterson map. Patterson maps are three-dimensional arrays calculated as the Fourier transform of

intensities obtained from diffraction patterns. The information in this data is dense, typically with many overlapping peaks, making it difficult for a human to interpret. USolv utilises modified versions of the Vit-V-Net and U-Net neural network architectures to extract information in the maps for estimating solvent content. Using a Python interface, the model can be run using .cif, .map, or .mtz files, in addition to a .txt file containing lists of these file types.

The Future

A PhD project has now been designed that incorporates improving the initial USolv model by working with additional information. This could then be extended to other properties, such as identifying structural features. In addition to predicting solvent content, USolv embeds the Patterson map of a given structure in a vector representation, as part of the neural network architecture. This vector space shows potentially useful clustering of similar structures that may have applications in areas such as identifying contaminants.

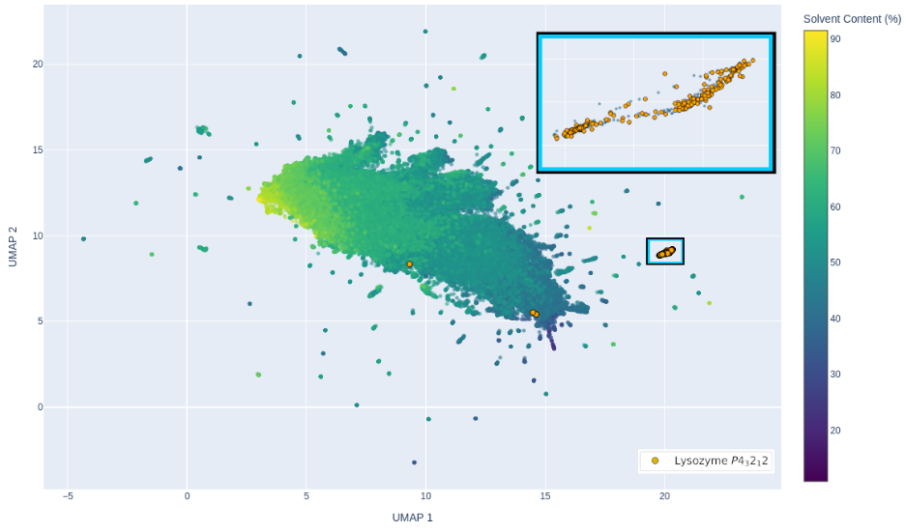

Figure: Example clustering of protein crystals in the Protein Data Bank (PDB). Structures highlighted in orange are different crystals containing lysozyme, an antimicrobial enzyme. The main cluster are all structures containing lysozyme from hen egg, whereas the three structures separated from the main cluster were found to be a lysozyme in complex with another molecule, and lysozyme obtained from a different source (Meretrix lusoria), containing a different amino acid sequence.

LICENCE	Programming Language	Repository URL
MIT	Python	https://github.com/ccp4/predict_solvent_content